

# Contextualized Language Model-based Named Entity Recognition in Slovak Texts

supervisor: Mgr. Endre Hamerlik  
consultant: Mgr. Marek Šuppa  
student: Bc. Dávid Šuba

## ● Named Entity Recognition (NER)

- recognition and categorization of named entities in text
- Person, Location, Organization, Product, Time .....
- one of the base tasks in NLP (applications: information extraction from CV, recommender systems...)

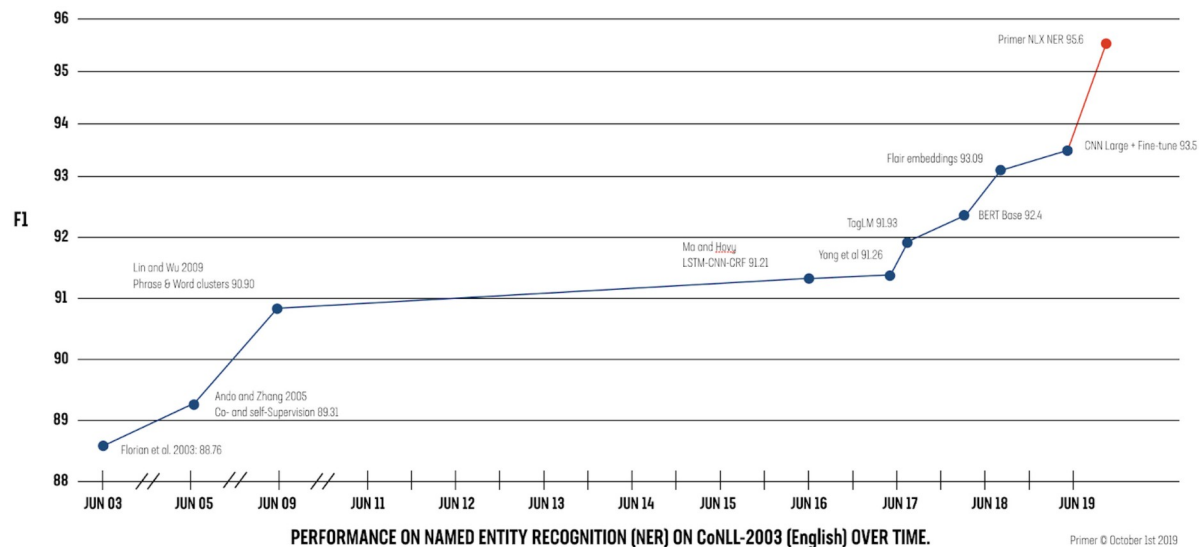
### NER DEFINITION

Luke Rawlence PERSON joined Aiimi ORG as a data scientist in Milton Keynes PLACE, after finishing his computer science degree at the University of Lincoln. ORG

source: <https://www.aiimi.com/insights/aiimi-labs-on-named-entity-recognition>

## Motivation

- state-of-the-art are contextualized language models based on deep neural networks
- the best systems for NER in slovak texts are rule based and are far behind [1]
- the main reason is lack of resources and datasets



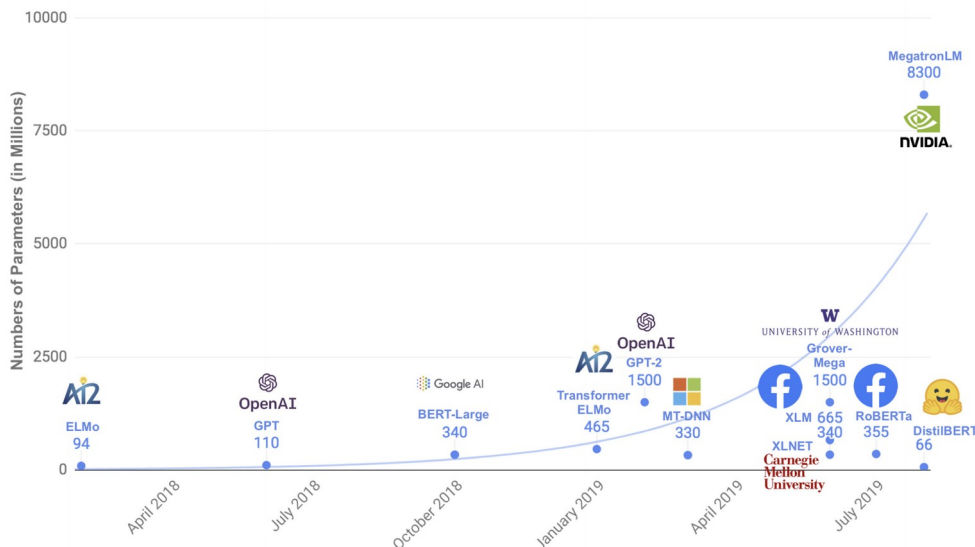
source: <https://company.primer.ai/blog/a-new-state-of-the-art-for-named-entity-recognition/www.aiimi.com/insights/aiimi-labs-on-named-entity-recognition>

## ● Goals

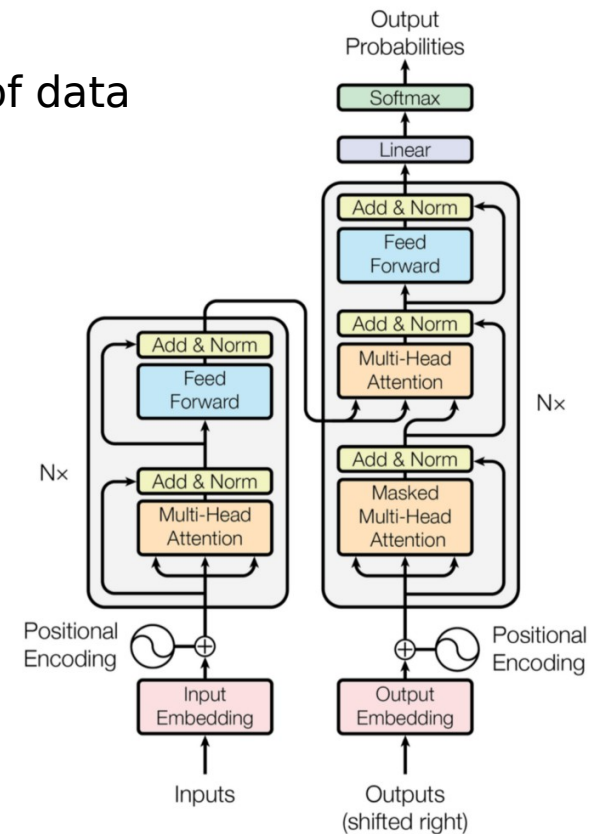
- utilize, improve and expand existing slovak NER datasets
- use large multilingual language models and transfer learning for named entity recognition in slovak texts

# Transformers language models

- BERT (Google), GPT-2/3 (OpenAI), RoBERTa (Facebook), ALBERT (Google)
- deep-learning models with attention layers
- state-of-the-art standard
- trained unsupervised on huge amounts of data
- fine-tuning for specific tasks (transfer learning)



source: <https://venturebeat.com/2020/02/10/microsoft-trains-worlds-largest-transformer-language-model/>



source: [2]

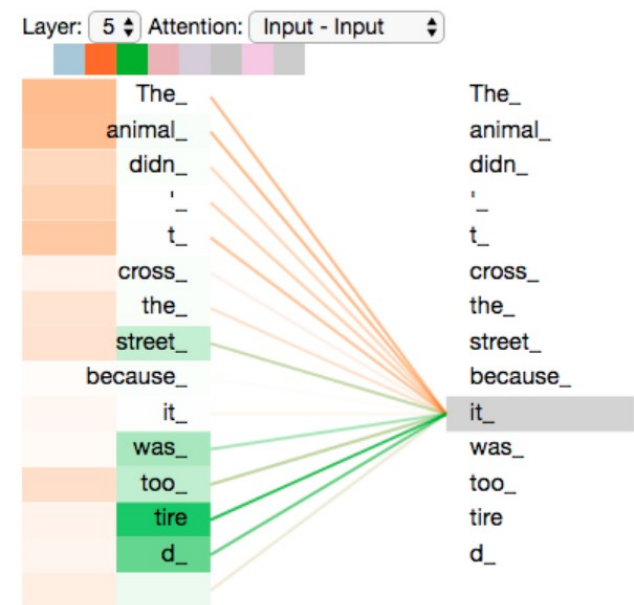
## Transformers language models

- attention: the attention-mechanism allows model to decide at each step of sequence which other parts of the sequence are important
- Q - query vector (word in a sequence)
- K - keys (all words in sequence)
- V - values (all words in sequence)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- positional encoding: allows model to utilize order of sequence

$$PE_{(pos, 2i)} = \sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos, 2i+1)} = \cos(pos/10000^{2i/d_{\text{model}}})$$



source: <http://jalammar.github.io/illustrated-transformer/>

## ● Silver Standard Dataset (WikiANN)

	Precision	Recall	F1
<b>Trankit</b>	88.9	88.51	88.7
<b>Spacy:</b> tok2vec + Transition-based parsing	84.49	82.96	83.72
<b>Spacy:</b> Multilingual Bert + Transition based parsing	92.64	92.33	92.49

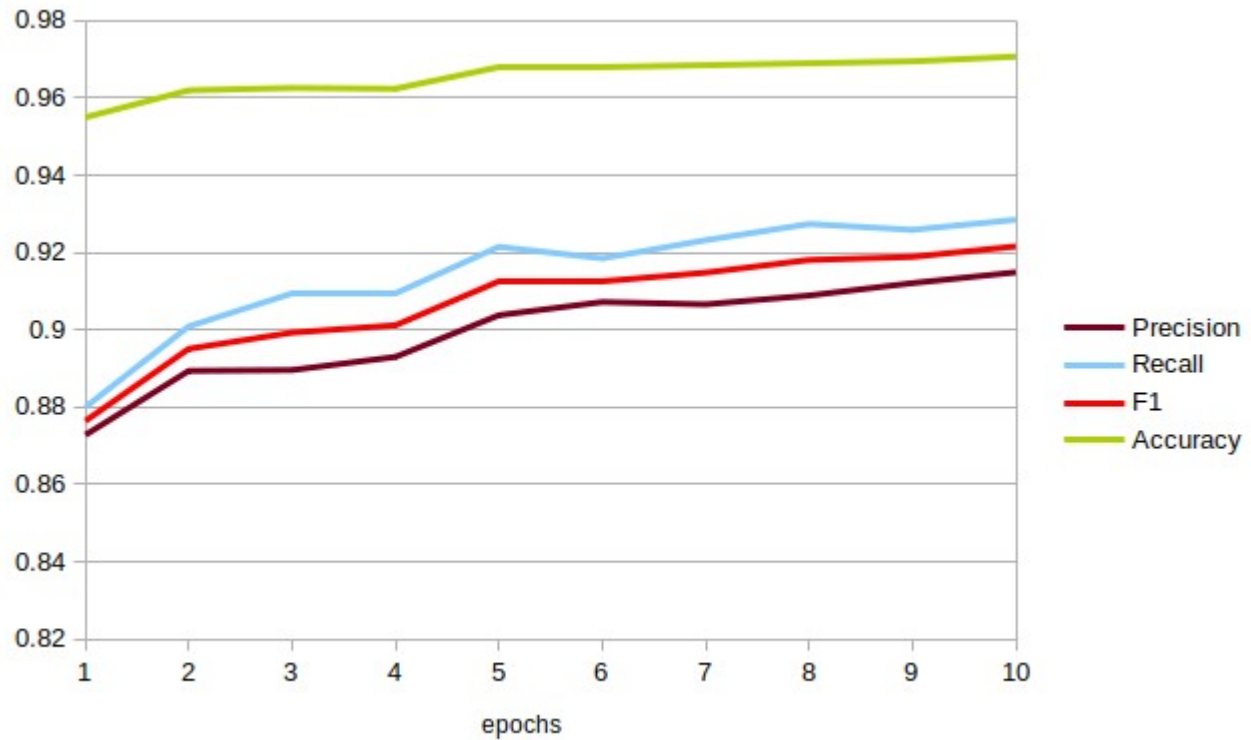
## ● Silver Standard Dataset (WikiANN)

- Oct 2021: **SlovakBERT** [1]
- KIIT and Gerulata
- the first Slovak-only transformers-based model trained on a sizeable corpus
- RoBERTa architecture, Web-crawled corpus (19.35 GB)
- evaluated on multiple NLP downstream tasks (however not NER) and achieved state-of-the-art results
- we finetuned it with WikiANN dataset utilizing HuggingFace NER pipeline [2] with following results:

Precision	Recall	F1
91.4	92.8	92.1



## ● Finetuning SlovakBert model

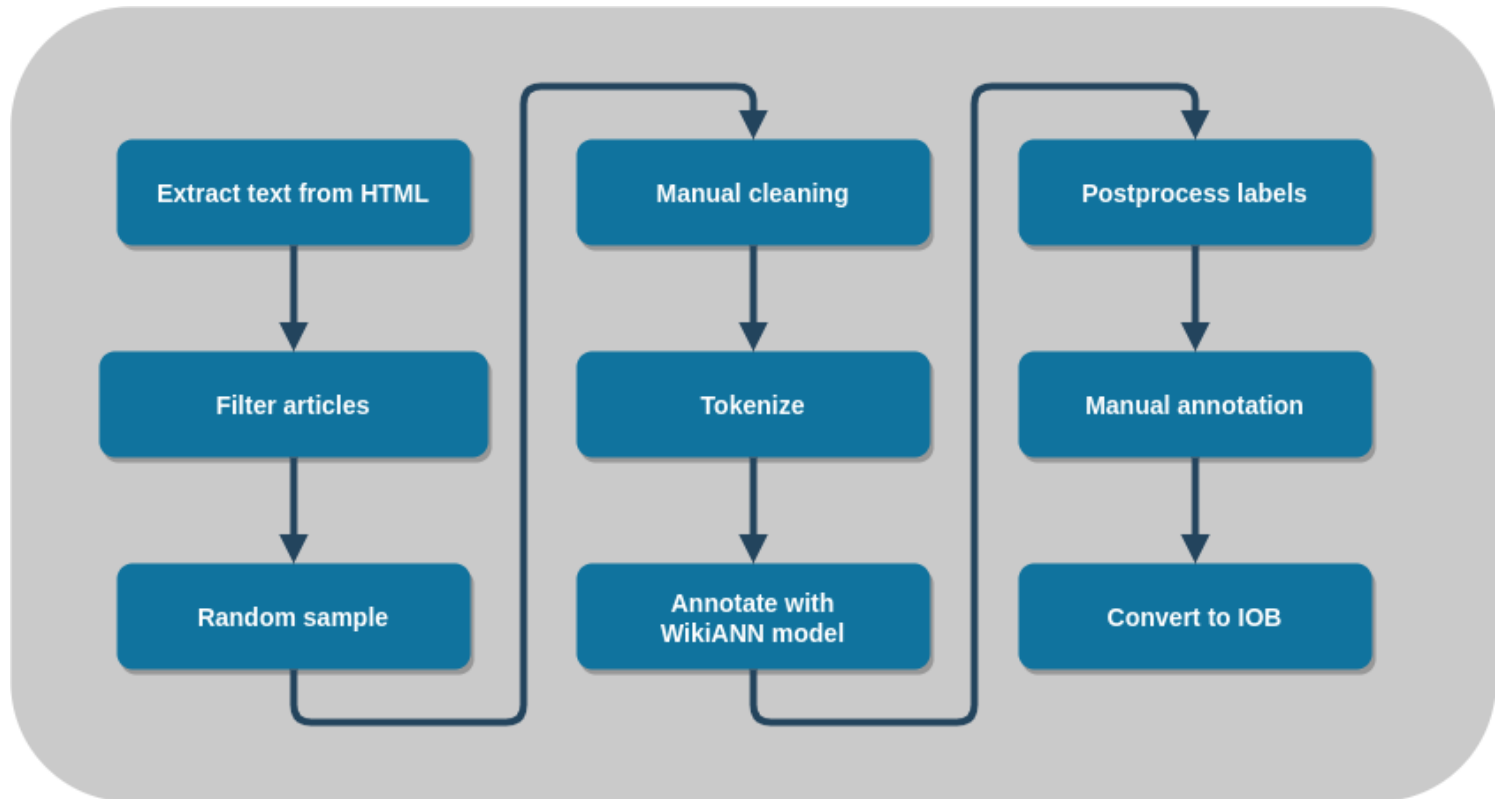


## ● Gold Standard dataset creation

- necessity for proper evaluation
- possible sources:
  - Slovenský Národný Korpus [2]
  - newspaper web pages (SME, Aktuality, ...)
  - legal texts (slov-lex.sk, ru.justice.sk)
  - Wikipedia
- license without limitations is priority
- need for entity-rich texts
- only Wikipedia has Creative Commons license
- [dumps.wikimedia.org/skwiki/latest/](https://dumps.wikimedia.org/skwiki/latest/)

## ● Gold Standard dataset creation

- Goal: 15k labeled entities



# Prodigy

- before manual labeling

**prodigy**

**PROJECT INFO**

DATASET

test06

LANGUAGE

en

RECIPE

ner.manual

VIEW ID

ner\_manual

**PROGRESS**

THIS SESSION

3

TOTAL

9

ACCEPT

3

REJECT

0

IGNORE

0**HISTORY**

Murad II . alebo Murat II . ( ara... ✓

Andrej Šeban ( \* 23. jún 1962 , ... ✓

Vláda je najvyšším ( spravidla ... ✓

© 2017-2022 Explosion (Prodigy v1.10.8)

PER 1

ORG 2

LOC 3

MISC 4

MUDr. Peter Osuský PER , CSc. PER ( 11. október 1953 , Bratislava LOC ) je poslanec Národnej rady Slovenskej republiky ORG , poslanec miestneho zastupiteľstva bratislavskej mestskej časti Staré Mesto LOC , vysokoškolský pedagóg a lekár , bývalý prodekan Lekárskej fakulty Univerzity Komenského ORG , a bývalý prorektor Univerzity Komenského v Bratislave ORG .

Je spoluautorom viacerých odborných monografií a člen autorského kolektívu Encyklopédia Beliana ORG , nositeľ Zlatej Jánskeho PER plakety .

Je členom Slovenskej olympijskej akadémie ORG a Slovenského olympijského výboru ORG .

V rokoch 1991-1995 bol predsedom Zväzu slovenských filatelistov ORG .

Peter Osuský PER je ženatý a má 4 deti .

Peter Osuský PER je spoluautorom viacerých odborných monografií a členom autorského kolektívu Encyclopaedia Beliana ORG .

ID: 41027



# Prodigy

- after manual labeling

**prodigy**

**PROJECT INFO**

DATASET

test06

LANGUAGE

en

RECIPE

ner.manual

VIEW ID

ner\_manual

**PROGRESS**

THIS SESSION

3

TOTAL

9

ACCEPT

3

REJECT

0

IGNORE

0

**HISTORY**

Murad II . alebo Murat II . ( ara...

✓

Andrej Šeban ( \* 23. jún 1962 , ...

✓

Vláda je najvyšším ( spravidla ...

✓

© 2017-2022 Explosion (Prodigy v1.10.8)

PER 1 ORG 2 LOC 3 MISC 4

MUDr. Peter Osuský PER , CSc. ( 11. október 1953 , Bratislava LOC ) je poslanec Národnej rady Slovenskej republiky ORG , poslanec miestneho zastupiteľstva bratislavskej mestskej časti Staré Mesto LOC , vysokoškolský pedagóg a lekár , bývalý prodekan Lekárskej fakulty Univerzity Komenského ORG a bývalý prorektor Univerzity Komenského v Bratislave ORG .

Je spoluautorom viacerých odborných monografií a člen autorského kolektívu Encyklopédia Beliana MISC , nositeľ Zlatej Jánskeho plakety MISC .

Je členom Slovenskej olympijskej akadémie ORG a Slovenského olympijského výboru ORG .

V rokoch 1991-1995 bol predsedom Zväzu slovenských filatelistov ORG .

Peter Osuský PER je ženatý a má 4 deti .

Peter Osuský PER je spoluautorom viacerých odborných monografií a členom autorského kolektívu Encyclopaedia Beliana MISC .

ID: 41027



## ● Next steps

- train/finetune most common NER models with created dataset
- use data augmentation method
  - Pattern Exploiting Training [3] – uses masked language modeling with pretrained language models for few-shot learning
    - ‘I am student of Matfyz. Matfyz is <masked>.’
- utilize transfer learning from close languages (e.g. czech)

## ● Literatúra

- [1] Kaššák, Ondrej - Kompan, Michal - Bieliková, Mária. Extrakcia pomenovaných entít pre slovenský jazyk. In ZNALOSTI 2012 : Sborník příspěvků 11 ročníku konference, Mikulov, hotel Eliška 14.-16. 10. 2012. Praha:Matfyzpress, 2012, pp. 52–61
- [2] Pikuliak, Matúš, et al. SlovakBERT: Slovak Masked Language Model. arXiv preprint arXiv:2109.15254 (2021).
- [3] <https://github.com/huggingface/transformers>
- [4] <https://korpus.sk/>
- [5] Tam, Derek, et al. Improving and simplifying pattern exploiting training. arXiv preprint arXiv:2103.11955 (2021)

## ● Key articles

- Suppa, Marek and Jariabka, Ondrej. 2021. Benchmarking Pre-trained Language Models for Multilingual NER: TraSpaS at the BSNLP2021 Shared Task. In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, pages 105–114, Kiyv, Ukraine. Association for Computational Linguistics.
- Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for ner. arXiv:1902.00193
- Pikuliak, Matúš, et al. SlovakBERT: Slovak Masked Language Model. arXiv preprint arXiv:2109.15254 (2021).





Thank you for your  
attention!